Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds.

Joanne Cleland, Caitlin McCron & James M. Scobbie.

Queen Margaret University

Abstract

Analogous to lip-reading, there may be a natural human capacity to 'tongue-read'. Although the ability of untrained participants to perceive aspects of the speech signal has been explored for some visual representations of the vocal tract (i.e. talking heads), it is not yet known to what extent there is a natural ability to interpret speech information presented through two clinical phonetic tools: EPG and ultrasound. This study aimed to determine whether there is any intuitive ability to interpret the images produced by these systems.

Twenty adults viewed real-time and slow motion EPG and ultrasound silent movies of 10 different linguo-palatal consonants and four vowels. Participants selected which segment they perceived from four forced-choice options.

Overall participants scored above chance in the EPG and ultrasound conditions, suggesting that these images can be interpreted intuitively to some degree. This was the case for consonants in both conditions and for vowels in the EPG condition.

Keywords

Electropalatography, ultrasound, speech-reading, visual feedback.

Introduction

It is well known that being able to view the face of a speaker enhances the intelligibility of an utterance by virtue of lip-reading (Benoit and Le Goff, 1998). A small number of studies have looked at whether a natural capacity might also exist for "tongue-reading". The tongue is a major articulator, involved in the production of most English consonants and all vowels, yet it is highly inaccessible and listeners are able to view it only partially at best. Despite this, listeners can copy someone else's speech characteristics from sound alone easily. Acquiring speech from sound exposure is a natural process, even for blind individuals. According to the Motor Theory of Speech Perception, listeners perceive speech sounds as the "intended phonetic gestures of the speaker" (Liberman and Mattingly, 1985), meaning listeners use articulatory knowledge, albeit at a subconscious level. Evidence for Motor Theory has been mixed; however the discovery of mirror neurones, or specifically echo neurones, has reignited interest in this theory (Lotto, Hickok and Holt, 2008). There is a vast literature supporting the view that mirror neurones are responsible for the imitation system, which may be the root of learning, but most of this literature investigates the visual domain in non-speech tasks. Evidence now exists that auditory perception of a sound (and theoretically a speech sound) is directly related to the action required to make that sound. That is, upon hearing a sound, echo neurones responsible for the action required to generate that sound will fire (Kohler, Keysers, Umilta, Fogassi, Gallese, and Rizzolatti, 2002), so in primates the noise of a stick dropping will fire the neurone involved in the actual action.

It is hypothesised that the this echo neurone system could be essential in learning to speak. It therefore seems possible that typical listeners/speakers have access to the articulatory information involved in speech production and would be able to make use of visual information about normally invisible articulators to enhance perception of speech, and perhaps even to learn new speech sounds. In primates the area of the brain containing echo neurones is analogous to Broca's area in humans (Kohler et al. 2002), providing further support of the role of echo neurones in

speech perception and production. While it is clear that a speech perception/ production link must exist, it is far from clear whether listeners have access to the articulatory information of speakers. That is, just because a listener understands a speaker, that does not mean the listener knows what the speaker's tongue (and other parts of the vocal tract) are doing during speech.

A small number of studies have attempted to assess whether listeners have an intuitive tongue-reading ability, using various "Talking Heads". Talking heads are artificial animations of speech production usually based on instrumental data of real speech (often Magnetic Resonance Imaging or electromagnetic articulograph). Some are 3D (e.g. Badin and Serrurier, 2006) and some are 2D (e.g. Kröger, 2003), but most attempt to model the movement of the tongue during speech by providing the user with a cut-away mid-sagittal view of the tongue (as in figure 1). The main application of Talking Heads is usually as a teaching tool for pronunciation training in second language learning, however, there is little evidence that this is effective.

However, there is increasing evidence that listeners are able to use information about the tongue to enhance perception of native speech sounds. Badin, Tarabalka, Elisei, and Bailly (2010) investigated the ability of listener-viewers to use a Talking Head to enhance perception of speech in various noise conditions. The mean phoneme identification rate was significantly greater for all conditions where audio-visual information was added (including a lip-only condition), but more importantly phoneme recognition was significantly greater (68.1%) when a mid-sagittal view with the tongue visible was compared to a mid-sagittal view with no visible tongue (63.7%). Badin et al. (2010) concludes that there is a natural, intuitive, capacity for listeners/viewers to tongue-read. This provides support for a perception/production link which could relate to the theory of mirror neurones. In a different study, Kröger, Gotto, Albert, Neuschaefer-Rube, C (2008) show that children as young as 4;6 with articulation disorders show a similar ability. They tested phoneme recognition of silent animations (based on MRI with information about all articulators) whereby the children were asked to watch the animations and produce what they thought the speech sound was.

Responses were rated on a scale for phonological feature correctness. There appears to be a confound here, with children potentially asked to produce sounds that were not in their phonetic inventories (since they had articulation disorders), but this is not explored in the paper. Since no child achieved 100% it is possible that children were unable to correctly produce the speech sounds that were usually in error in their speech.

Nevertheless, these children with articulation disorders do show some ability to tongue-read. it does not, however, necessarily follow that Talking Heads are a useful tool for teaching children the speech sounds they have failed to acquire naturally. Only one study has attempted to investigate this issue. Fagel and Madany (2008) use a Talking Head to treat interdentalised /s/ and /z/ in German children. Six out of eight children lessened their degree of lisping after just one learning session but the authors were unable to demonstrate that the improvement was a direct result of the Talking Head intervention. Most speech therapists, especially in the UK, will question the necessity of a mid-sagittal Talking Head for remediation of interdental sibilants since the incorrect production would be easily viewed on the face of another speaker- or in a mirror if visual feedback is required. However Talking Heads may be of some use for indirect approaches in therapy.

In clinical phonetics, researchers and clinicians will be more familiar with instrumental techniques which provide visual feedback of the speakers' own articulations. Electropalatography (EPG) for example is a technique for displaying the timing and location of tongue-palate contact (Hardcastle and Gibbon, 1997). The disordered speaker sees an abstract representation of linguo-palatal consonants (and some information for high vowels) in real time, and is encouraged to use this to modify their own erroneous articulations. Clinically, this computer-based therapy tool has been used widely to provide visual feedback to remediate speech sound disorders (Bernhardt Gick, Bacsfalvi and Ashdown, 2003) with positive results reported in a large number of case and small group studies. The understanding of this visual display is thought to be relatively intuitive (Gibbon

and Wood 2010), even for those with cognitive impairment (Cleland, Timmins, Wood, Hardcastle and Wishart, 2009). Although its therapeutic success has been reported there has been little exploration of why it might be useful for the speaker to view their own articulation and precisely how the presence of a real-time visual image of tongue-palate contact is able to help after disordered productions become habitualised.

There is no existing evidence supporting the hypothesis that there may be a natural capacity to interpret, or tongue-read, EPG. Clinical application of EPG usually follows training and demonstration, in conjunction with instruction-based direct therapy provided by a specialist Speech and Language Therapist (Gibbon and Wood 2010) and it is entirely possible that the power of EPG lies more in its diagnostic value (since it can be used to create a fine grained analysis of speech and to suggest underlying causes of speech disorders) than exploitation of the putative mirror neurone system whereby actually showing the child a correct articulation would lead to improvement.

EPG differs from Talking Heads in two important ways; firstly, the display is an abstract representation of one aspect of speech production, rather than an anatomically correct representation of a speaker's mouth. Secondly, it is almost exclusively used as a real-time feedback tool by an SLT, rather than a model alone.

No studies report on whether there is a capacity to tongue-read from pre-recorded EPG. This is both theoretically and clinically interesting. If there is a capacity to tongue-read from EPG then it might be used like a Talking Head, avoiding the need to make expensive palates for each speaker. Alternatively, if feedback of the speaker's own speech is required, then ability to tongue-read might speed up the therapeutic process, or make indirect therapy (where the child uses the EPG equipment exclusively at home) a viable approach. Furthermore, Badin et al. (2010) report that some participants in their study were "good tongue-readers" whilst others were "poor tongue-readers". Whilst it is possible that poor tongue-readers might be receptive to training, if this is not

the case then having such an evaluation before offering EPG therapy might be a useful way of screening out those who are not likely to benefit. Moreover, it might give some clues as to why not all children have benefited from EPG in the past.

Another visual feedback technique which is gaining popularity is ultrasound tongue imaging (UTI). With this technique most of the surface of the tongue can be made visible in a mid-sagittal view in real-time. This view can be used for visual feedback of tongue and interpreting such images is thought to be relatively intuitive (Bernhardt, Gick, Bacsfalvi, and Adler-Bock, 2005). Unlike EPG, the image is an anatomically correct representation of a slice of the tongue, as in a Talking Head. However, other relevant anatomical information, such as the lateral margins of the tongue in the sagittal plane and the relation of the tongue to the hard palate, are not visible in UTI. Also, during speech the tip of the tongue is may be in shadow from the speaker's jaw or invisible due to a sublingual airpocket. However, since the tongue itself is imaged, rather than tongue-palate contact, ultrasound shows fuller information for a variety of segments, especially perhaps for vowels. The viewer can see the shape and location of the tongue change from one sound to another, based on ultrasonic echoes from structures within the tongue, and, more obviously, from the tongue's surface. Figure 1 compares ultrasound and EPG with a typical mid-sagittal diagram of [t].

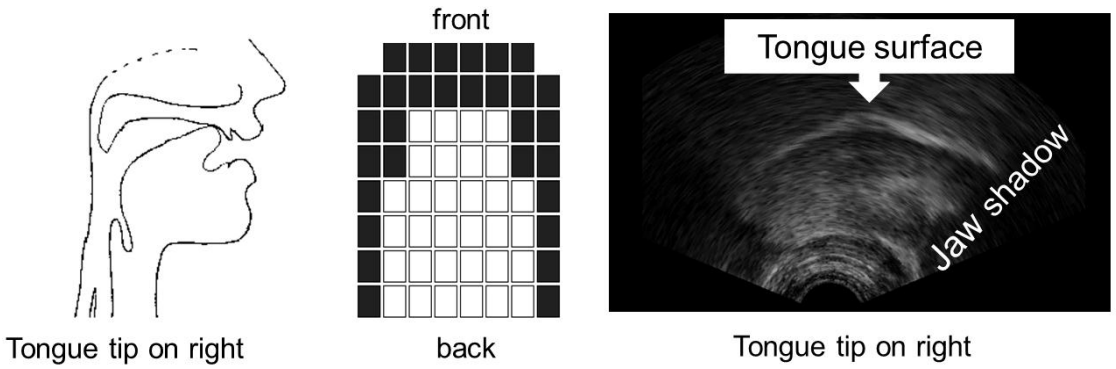

Figure 1: (left to right) comparison of typical mid-sagittal diagram, EPG, and ultrasound of [t].

Ultrasound might have an application as a Talking-Head-like model, but this has not been investigated. Models derived from ultrasound might have an advantage over models derived from EMA or MRI since data can be acquired quickly and easily at a high sample rate (Wrench and Scobbie, 2011) and from multiple speakers. The suitability of this technique for capturing the articulation of children is particularly useful since at present Talking Heads are based on adult speech. This might give a more realistic model, especially since children's speech is likely to differ from that of adults, while children are a key target group for speech therapy.

Aims

As a first step to determine whether tongue-reading can be observed with EPG and UTI, we sought to determine whether naïve adults, without disorders of speech, can identify a single segment from silent videos of EPG and/or ultrasound. The research questions were:

1. Is there a natural human capacity to tongue-read (a) EPG and (b) ultrasound displays?

2. Is the effect, if any, stronger with either technique?

3. Are vowels or consonants easier to tongue-read?

We predicted that, like the studies of tongue-reading with Talking Heads, there would be some capacity to identify segments above chance level in both techniques. Furthermore, we hypothesised that the percentage of lingual-palatal consonants correctly identified will be higher for EPG whilst vowels will be better discriminated in the ultrasound condition.

Method

Participants

Ten male and ten female typical Scottish-English speakers aged 20 to 22 (M=21, SD=0.71) were

recruited. Participants were excluded if they had any disorders of speech, or any related disorders, such as dyslexia. Participants were final year undergraduate students and none had previous experience viewing EPG or lingual ultrasound displays. None were students of linguistics, Speech and Language Therapy or related disciplines.

Stimuli

Simultaneous EPG and ultrasound video recordings of a female Scottish speaker were made using Articulate Assistant Advanced™ (AAA) (Articulate Instruments Ltd. 2011). Each target segment was recorded three times. Consonants were placed between open vowels to highlight the lingual gestures required. Vowels were prolonged. Table 1 shows the stimuli.

The consonants chosen were all present in, or specific to, Scottish English and allowed for a variety of place and manner of articulation. Only lingual consonants were selected, as these can be imaged using either EPG or ultrasound. Voicing was not assessed, as this cannot be observed using these visual feedback tools. It was not expected that participants would be able to intuitively distinguish between consonants sharing the same main place of articulation, for example [t] and [n]; these were therefore not offered alongside each other in the forced-choice task (see below). It was anticipated that participants would, however, observe a difference between consonants such as [t] and [tʃ], as dynamic information was available.

Table 1: Test segments

Consonants	**Vowels**
[ata]	[i:]
[ana]	[ɛ:]
[asa]	[a:]

[aʃa]	[ɔ:]
[atʃa]	
[aɹa]	
[aça]	
[aja]	
[aka]	
[axa]	

Training materials

To avoid the need for a large number of practice items, each participant was orientated to both the EPG and ultrasound displays using a scripted presentation with silent videos of practice segments. A tutorial was designed to briefly describe both EPG and ultrasound, demonstrating how to read each display without revealing any specific information regarding the test segments. [l] and [ŋ] were used as examples of 'front' and 'back' sounds, these segments were therefore not used in the main experiment.

Procedure

Each of the 14 test sounds were assessed four times resulting in 56 tokens per condition (EPG or UTI), 112 in total. The order of the tokens was randomised within each condition.

Following the tutorial participants individually viewed silent movies of each condition. Order of presentation (EPG or UTI first) was counterbalanced. Participants first viewed the test item in real-time and then in slow-motion (four times slower) and identified the segment from a four-option forced-choice. Segments were presented within words that demonstrated appropriate pronunciation. To illustrate, [ç] was presented as 'huge' and [x] as 'loch'. Example words were taken from Hewlett and Beck (2006, p.48). For each consonant, forced-choice options consisted of the

correct target segment and three distractor segments, one differing in place, one in manner (and perhaps place) and one non-lingual consonant. This meant there was only one possible correct answer for each test item. Since the four test vowels differed in tongue height and position they were all available for selection in the forced-choice options. Figure 2 shows an example test item. On completion of both conditions, participants were asked which style of visual feedback they preferred and why and gave a prediction as to which condition they performed better in. The full procedure took approximately 60 minutes.

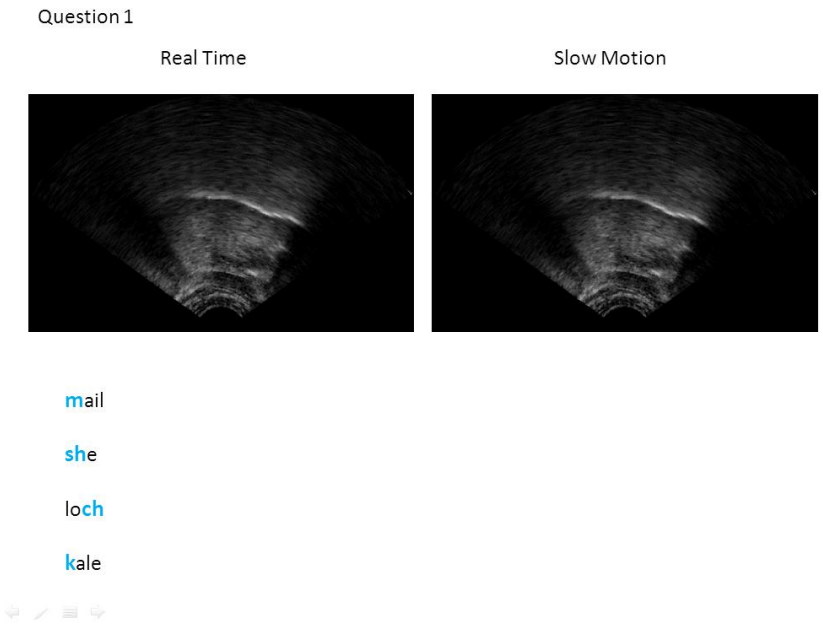

Figure 2: Example test item from the ultrasound condition. Videos were clickable.

Analysis

Two levels of analysis were carried out, broadly similar to Kröger et al. (2005). Firstly, a percentage segment correct criterion was applied. Since the forced-choice was carefully designed, it was possible for participants to score 100%. Secondly, a feature analysis was applied on a four-point scale. Correct selections received a score of three. Selection of segments produced at the same place of articulation, for example, an alveolar plosive for an alveolar fricative, received a score of two. Selections of segments produced in an adjacent place of articulation, for example a post-alveolar

fricative rather than an alveolar fricative, received a score of one. Since in UTI the hard palate is not visible (unless the speaker is swallowing or similar) it may be difficult to discern for example the difference between [s] and [ʃ]. All other selections received a score of zero. This four-point scoring method was not used for the vowel tokens, as they varied noticeably in tongue height and position. Percentage consonants correct results were compared to level of chance (25%), as in Kröger et al.'s (2008) study of the natural ability to interpret 2D and 3D models.

Results

Participants were able to identify which segment had been uttered from a silent movie of EPG 52% of the time (SD = 15.75) and from a silent movie of UTI 41% of the time (SD = 11.28). Like Badin et al. (2010), there were both good and poor tongue-readers with scores ranging from 18% to 82% for EPG and 23% to 61% for ultrasound. The majority of participants scored significantly above chance (Chi-square: EPG condition, (19, n = 20) = 403.24, $p < 0.001$; ultrasound condition, 2(19, n = 20) = 144.17, $p < 0.001$), showing that most people display a natural capacity to tongue-read from these techniques. Figure 3 shows the individual results for participants, with chance level (25%) indicated.

Overall there was a highly significant difference between correct answers achieved in the EPG and ultrasound conditions, $p < 0.001$, suggesting that EPG is more conducive to tongue-reading for the segments tested here. There was a strong correlation evident ($r = 0.55$, $p = 0.01$) between performance in the EPG and UTI conditions.

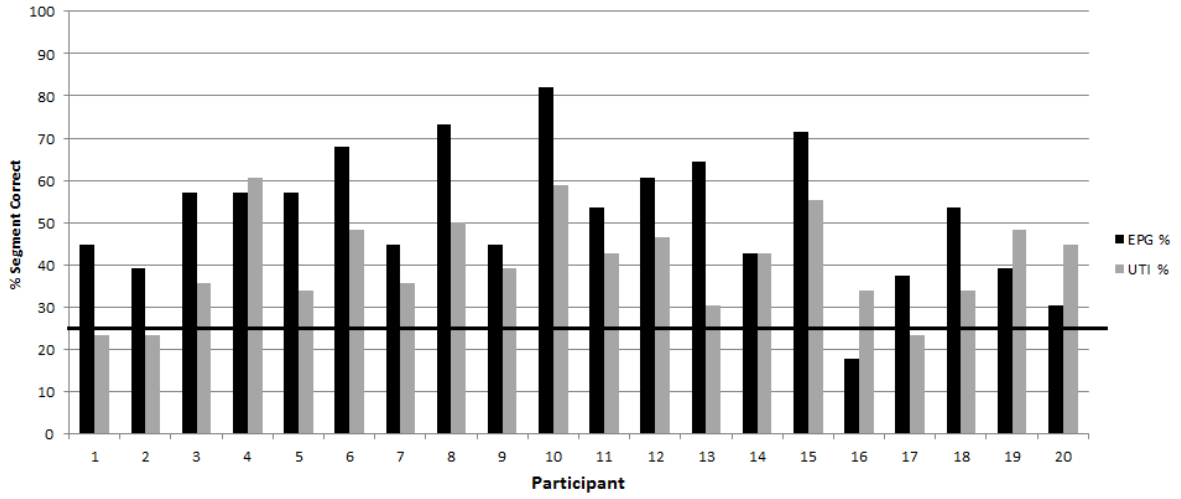

Figure 3: Individual results with chance (25%) represented by a line.

Comparing consonants and vowels

Figures 4a and 4b show the group results for consonants (4a) and vowels (4b) in each condition. Percentage consonants correct was 55% in the EPG condition (SD = 18.52) and 46% in the ultrasound condition (SD = 13.30). The large standard deviations (especially in the EPG condition) reflect the heterogeneity in individual performance. Again, in both conditions performance was above chance (Chi-Square: EPG condition, 2 (19, n = 20) = 387.21, $p < 0.001$; ultrasound condition, 2 (19, n = 20) = 192.67, $p< 0.001$). Consonants were more easily tongue-read with EPG than ultrasound ($p<0.001$).

For vowels, correct identification was lower: 44% for EPG (SD = 18.65) and only 26% for ultrasound (SD = 14.98). This was found to be at chance level for the ultrasound condition, (2 (19, n = 20) = 0.27, $p > 0.20$). Surprisingly, this was not the case for the EPG condition, where the percentage of correct vowels was significantly above chance level, 2(19, n = 20) = 62.02, $p < 0.001$. Contrary to our hypothesis, vowels were more easily tongue-read with EPG than ultrasound ($p<0.001$).

As the number of vowels and consonants tested were not equal, it is difficult to assess the difference in performance between these. However, this is assumed to be significant in the ultrasound condition, as the identification of consonants above chance level was found to be highly

significant whilst vowels were not. Participants appear to be more successful in tongue-reading consonants than vowels in both conditions.

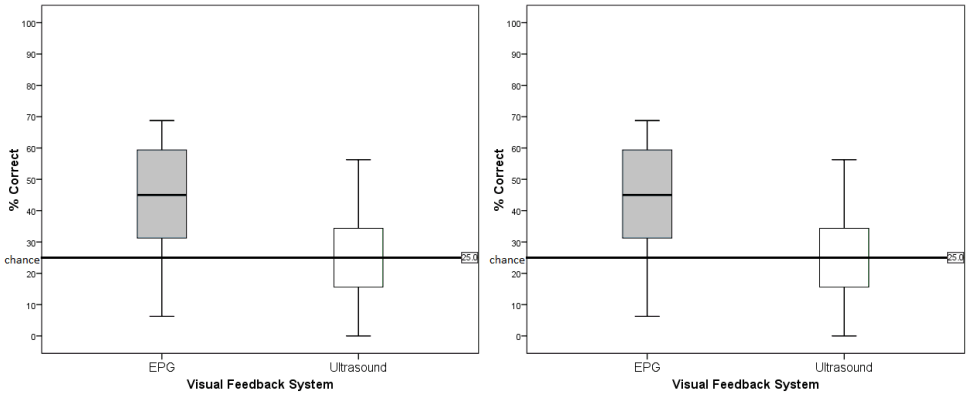

Figure 4a: Group results of consonants in EPG and ultrasound ; Figure 4b: Group results of vowels in EPG and ultrasound

Feature analysis

Table 2 compares the scores from the strict scoring criteria with the near-feature analysis. As expected, when a feature analysis was applied rather than strict right/wrong criteria the % correct increased, suggesting that participants sometimes made errors involving the same or adjacent place of articulation. As expected, the scores obtained with both methods was very strongly correlated: EPG condition, r = 0.98, p = <0 .001, ultrasound condition, r = 0.99, p = <0 .001.

Table 2: Comparison of 2-level scoring.

EPG Mean % Correct		UTI Mean % Correct	
Strict Correct Segment	4-Point Scale	Strict Correct Segment	4-Point Scale
51.96	58.07	40.54	46.31

Participants' preferences

In a qualitative debrief 60% of female participants reported a preference for EPG while 80% of males specified that they preferred EPG. Therefore, overall, 70% (14) of participants preferred the EPG display whilst the remaining 30% (6) reported that they found ultrasound easier to understand. It is interesting to note that participants' preferences were not always justified. Participants 4 and 16 preferred EPG, however they were more successful in the ultrasound condition. Participants 2, 9, 12 and 17 all preferred ultrasound yet performed better in the EPG condition (see figure 3).

Discussion

Previous research has found some natural capacity to tongue-read from mid-sagittal animations of the vocal tract, despite the fact that speakers will have little or no opportunity to observe tongue motions naturally. Our experiment extends this to instrumental methods commonly used in phonetic research and speech and language therapy.

Overall, consonants were easier to tongue-read than vowels, supporting the view of Speech and Language Therapists that EPG is most useful for remediation of consonant errors (Gibbon and Paterson, 2006). It was surprising that participants performed at chance level in the ultrasound vowel condition since previous research has highlighted the value of this visual feedback tool in treatment of vowels due to the anatomically correct visualisation of the configuration and position of the tongue (Bernhardt et al., 2005). Difficulty tongue-reading from vowels may be due to a speaker's lack of awareness of their own tongue during vowel production or because vowel quality is highly dependent on the shape and width of the whole vocal tract, not just tongue location and shape. Clinically, this might suggest that remediation of vowel disorders with ultrasound may be highly dependent on training from a specialist speech and language therapist. Theoretically, it is difficult to reconcile why participants had such difficulty with vowels if we subscribe to Motor Theory and/or mirror neurone theory, especially since most speakers acquire vowels easily and early in development. Other researchers have also suggested that mirror neurones do not play as central

a role in speech as first hypothesised. Motor Theory would predict that since there is a direct link between perception and production, damage to Broca's area (if it contains echo neurones in humans) would result in parallel difficulties in speech production and perception. Studies of people with lesions in this area do not support this (Lotto et al., 2008). If mirror neurones are not at play then perhaps our participants were using a much more conscious strategy to complete the tongue-reading task, perhaps watching the silent movie, then silently articulation each of the forced-choice options to find a match. It would be interesting to investigate this using a paradigm where the participants were recorded using ultrasound or EPG while they undertook the perception task.

Despite ultrasound showing an anatomically correct representation of the central tongue slice similar to Talking Heads, and despite EPG using an abstract representation, participants were more successful at tongue-reading from EPG. It is known that speakers make use of tactile feedback provided by tongue-palate contact in order to detect lingual position and movement in consonant production (Hewlett and Beck, 2006). This may explain why participants had more success in intuitively reading these silent videos, as EPG provides a visual representation of a tactile event, tongue-palate contact. Moreover, the EPG display is normalised across speakers, perhaps making it easier to tongue-read when, as in this experiment, viewing the tongue movements of an unfamiliar speaker. In contrast, ultrasound is individualised for each speaker, therefore an experiment which tests how well a speaker can interpret their own pre-recorded ultrasound tongue movements may have been more successful.

We asked participants which instrumental method they preferred and most (70%) had a preference for EPG. Qualitatively, participants commented on the benefit of the precise contact points and enjoyed the layout of the EPG display. Some said they found the mid-sagittal view provided by the ultrasound display confusing. They reported that they could ascertain which patterns would be produced by each sound and found it easier to locate the place of articulation using EPG. Again, these comments support the idea that some participants may have been using a

strategy to complete the task, rather than unconsciously making use of a mirror neurone system. Participants also appreciated EPG's use of colour, despite this being arbitrary. Even those that reportedly preferred ultrasound often described this feedback tool as 'unclear' or 'fuzzy'. However, participants did not mention any negative impact caused by the shadowed tip of the tongue in ultrasound. Those that preferred ultrasound reported that they benefitted from viewing an 'actual tongue'. These participants felt that this made it easier to appreciate the range and duration of movement.

Since EPG has a clear advantage over ultrasound, it may have some potential as a Talking-Head-like model where pre-recorded EPG is used to demonstrate speech sounds to either second language learners or people with speech sound disorders. The possible advantage of this over existing Talking Heads is that normative data exist for a small number of children (Timmins Hardcastle, Wood, and Cleland, 2011) and many more adults (e.g. McLeod and Roberts, 2005). It is also relatively straightforward to average data across speakers, since the EPG display is already normalised. However, if visual feedback is required then ultrasound should still be considered as it is cheaper and more flexible than EPG since speakers do not require a custom-made artificial palate. Moreover, a current research project, Ultrax (2011), aims to make the ultrasound image more accessible by adding anatomical information, essentially making it more like a Talking Head and also allowing speakers to view tongue-palate contact. It is possible that this would enhance the tongue-reading potential of ultrasound.

Is tongue-reading essential for visual feedback success?

Although tongue-reading appears to be possible with both Talking Heads and instrumental phonetic techniques it is unclear whether it is a necessary step in using EPG or ultrasound for visual feedback. With both these techniques, therapy will usually involve either demonstration of the speech sound to be taught by the Speech and Language Therapist and/or drawing the speaker's attention to a

static target pattern. However, therapy mostly focuses on directed feedback, with the therapist acting as a crucial mediator, interpreting articulatory information and then instructing the speaker how to move their tongue in order to achieve the correct articulation. It is therefore possible that tongue-reading by clients may not be essential for these techniques to be useful, suggesting that even those who are "poor tongue-readers" (Badin et al., 2010), performing at chance, will still benefit from visual feedback therapy.

Studies which investigate the use of an articulatory model only to teach new speech sounds are few. Massaro, Bigler, Chen, Perlman and Ouni (2008) used a Talking Head to teach native English speakers a new vowel [y] and a new consonant [q]. While the view of the lips was successful for teaching the high-front rounded vowel [y], learners who had access to a mid-sagittal Talking Head for learning a contrast between /k/ and a uvular stop [q] had no advantage over those who used audio only. Similarly, the study by Fagel and Madany (2008) which used a Talking Head to teach [s] and [z] to children with interdental lisps was unable to show an effect. This perhaps suggests that a visual model is not sufficient for success. However, since the above studies did not give the learners any feedback (e.g. a Speech and Language Therapist telling the learner how close their production was to the target) a further study is required that compares an articulatory model with a visual biofeedback system using both the same type of display and with the same amount of support from a Speech and Language Therapist.

Summary and Conclusions

This study sought to establish whether naïve participants are able to determine which speech sound is being produced in silent videos of the dynamic aspects of speech production, using EPG and ultrasound. Most participants performed above chance, confirming some capacity for tongue-reading. It is still unknown how participants completed the task. How did they know which speech

sound the speaker was making? While there is most certainly some kind of perception-production link, our experiment does not offer explicit evidence for Motor Theory or for mirror neurones, since it was probably possible to complete the task offline by silently articulating each of the forced choice answers to find a plausible match for the articulation shown in the silent movie. This would also account for the fact that no participant achieved a ceiling score and some achieved a floor score, despite no history of any difficulty in learning to speak.

In sum, our findings support the notion that EPG an ultrasound are relatively intuitive techniques (Gibbon and Wood, 2010 and Bernhardt et al., 2005). Both techniques seem suitable for indirect therapy, since little training would be required in helping those with speech disorders to interpret the images. The ability to tongue-read from EPG and ultrasound varied hugely among participants with some individuals performing below chance level. However, most participants were able to tongue-read, perhaps giving some clues as to the mechanisms that underlie the success of EPG and ultrasound as therapeutic tools. The leap between tongue-reading native phonemes and using the displays to learn speech sounds which are not in the speaker's phonetic inventories still needs further investigation.

Acknowledgments

We wish to thank all the participants who gave up their time to complete the experiment and Xxxxx Xxxxxx for technical assistance.

Declarations of Interest

The first author is supported by an EPSRC grant (Grant no:xxxxxxxx).

References

Articulate Instruments Ltd. (2011). *Articulate Assistant Advanced User Guide: Version 2.13*. Edinburgh: Articulate Instruments Ltd.

Badin, P. & Serrurier, A. (2006). Three-dimensional linear modelling of tongue: Articulatory data and models. In H.C. Yehia, D.Demoline, R. Laboissiére,R. (Eds), *Seventh International Seminar on Speech Production*, *ISSP7*. Ubatuba, SP, Brazil, UFMG, Belo Horizonte, Brazil, pp395-402.

Badin, P., Tarabalka,Y., Elisei,F. & Bailly,G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, 52, 493-503.

Benoît, C., Le Goff, B. (1998). Audio-visual speech synthesis from French test: eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26, 117-129.

Bernhardt, B., Gick, B., Bacsfalvi, P. & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics and Phonetics*, 19, 605 – 617.

Bernhardt, B., Gick, B., Bacsfalvi, P. & Ashdown, J. (2003). Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners. *Clinical Linguistics and Phonetics*, 17, 199 – 216.

Cleland, J., Timmins, C., Wood, S.E., Hardcastle, W. J. & Wishart, J.G. (2009). Electropalatographic therapy for children and young people with Down's syndrome. *Clinical Linguistics and Phonetics*, 23, 926 – 939.

Fagel, S. & Madany, K. (2008). A 3-D virtual head as a tool for speech therapy for children. In: *Interspeech 2008*. Brisbane, Australia, 2643–2646.

Gibbon, F. and Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy*, 23 275 – 292.

Gibbon, F.E. and Wood, S.E. (2010). Visual Feedback Therapy with Electropalatography. In: A.L. Williams, S. McLeod and R.J. McCauley (Eds). *Interventions for Speech Sound Disorders in Children.* Baltimore: Paul H. Brookes Pub, 509 – 532.

Hardcastle, W., and Gibbon, F. (1997). Electropalatography and its clinical applications. In M. Ball, and C. Code (Eds.), *Instrumental Clinical Phonetics*, pp149–193, London: Whurr.

Hewlett, N. and Beck, J. (2006). *An Introduction to the Science of Phonetics.* London: Lawerence Erlbaum, pp.48.

Kohler E., Keysers C., Umilta A., Fogassi, L., Gallese, V. and Rizzolatti, G. (2002). Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons. *Science*,297, 5582, 846-848.

Kröger, B. (2003). Ein visuelles Modell der Artikulation. *Laryngo-Rhino-Otologie* 82, 402-407.

Kröger, B., Gotto, J., Albert, S. and Neuschaefer-Rube, C. (2005). A visual articulatory model and its application to therapy of speech disorders: a pilot study. *ZAS Papers in Linguistics* 40, 79-94.

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.

Lotto, A., Hickok, G. & Holt, L. (2008). Reflections on mirror neurones and speech perception. *Trends in Cognitive Sciences*, 13, 110-114.

McLeod, S.,& Roberts, A. (2005). Templates of tongue/palate contact for speech sound intervention. InC. Heine, & L. Brown (Eds.), Proceedings of the 2005 Speech Pathology Australia National Conference (pp. 104–112). Melbourne:Speech Pathology Australia.

Timmins, C., Hardcastle, W.J., Wood, S. and Cleland, J. (2011). An EPG analysis of /t/ in young people with Down's syndrome. *Clinical Linguistics and Phonetics*, 25, 1022-1027.

Ultrax. 2011. Overview of Research for Ultrax [online] Available at: http://www.ultrax-speech.org/research [Accessed November 17 2011].

Wrench, A. and Scobbie, J.M. (2011). Very High Frame Rate Ultrasound Tongue Imaging. *Proceedings of the 9th International Seminar on Speech Production (ISSP),* Canada, pp155-162.